PERBANDINGAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBORS UNTUK KLASIFIKASI TOPIK BERITA PADA SITUS DETIK.COM

Muhammad Iksan Maulana [1], Martanto [2], Umi Hayati [3]
[1,2] Teknik Informatika, STMIK IKMI Cirebon
[3] Sistem Informasi, STMIK IKMI Cirebon
Jl. Perjuangan No.10B, Karyamulya, Kec. Kesambi, Kota Cirebon
alqarniihsan22@gmail.com

ABSTRAK

Perkembangan pesat dalam bidang Informatika telah menjadi pendorong utama perubahan dalam berbagai aspek kehidupan manusia. Era digital saat ini menyaksikan revolusi teknologi yang telah mengubah cara kita berkomunikasi, bekerja, belajar, dan berbisnis. Studi-studi sebelumnya telah banyak mengkaji masalah klasifikasi berita, namun masih ada ruang untuk penelitian lebih lanjut. Algoritma Naïve Bayes Classifier yang digunakan sebagai metode untuk melakukan klasifikasi data terutama untuk kebutuhan deteksi terhadap berita berita palsu/fakta.Tujuan utama dari penelitian ini adalah untuk menguji dan membandingkan efektivitas Algoritma Naïve Bayes Classification dan K-Nearest Neighbors dalam klasifikasi topik berita. Dengan menggunakan Kedua nya bisa memberikan perspektip yang berbeda dan memungkinkan untuk mengeksplorasi kelebihan dan kekurangan masing masing metode, Dengan demikian kedua metode tersebut dipilih. penelitian ini akan memberikan pemahaman yang lebih baik tentang kemungkinan penerapan kedua metode ini dalam konteks yang relevan. Penerapan Algoritma Naïve Bayes Classification dan K-Nearest Neighbors dalam konteks klasifikasi topik berita berjumlah 9 tahapan. Tahapan tersebut adalah (1) Studi Literatur;(2) Pengumpulan Data;(3) Preprocessing Data;(4) Ekstraksi Fitur;(5) Pembagian Data;(6) Penerapan Algoritma Naïve Bayes Classification;(7) Penerapan K-Nearest Neighbors (K-NN);(8) Evaluasi Model;(9) Analisis Hasil. Dari hasil pengujian ini dapat di simpulkan bahwa pengujian analisis algoritma Naïve Bayes dan K-Nearest Neighbor ini didapatkan hasil akurasi terbaik dalam klasifikasi data yaitu pada algoritma K-Nearest Neighbor dimana mendapatkan akurasi sebesar 71,00% yang didapatkan pada data uji 10% dengan k = 1 dibandingkan algoritma Naïve Bayes yang hanya mendapatkan akurasi sebesar 66,00% pada data uji 10%. Perbandingan ini menunjukkan bahwa metode klasifikasi Naive Bayes dan K-Nearest Neighbors masing-masing memiliki kelebihan dan kekurangan dalam klasifikasi topik berita di Detik.com.

Kata kunci : *Klasifikasi, Berita, Naïve bayes, K-NN, Machine Learning*

1. PENDAHULUAN

Perkembangan pesat dalam bidang Informatika telah menjadi pendorong utama perubahan dalam berbagai aspek kehidupan manusia. Era digital saat ini menyaksikan revolusi teknologi yang telah mengubah cara kita berkomunikasi, bekerja, belajar, dan berbisnis. Perkembangan ini menciptakan tantangan dan peluang yang signifikan di berbagai bidang, termasuk teknologi informasi, ekonomi digital, pendidikan online, dan media massa. Teknologi informasi dan komputasi telah menjadi tulang punggung masyarakat modern, mengubah fundamental cara kita mendapatkan, mengolah, dan menyebarkan informasi. Dalam konteks ini, klasifikasi topik berita menjadi sebuah isu krusial, karena melibatkan kemampuan untuk menyortir dan mengorganisasi informasi yang sangat berlimpah di era digital ini.

Dalam era di mana akses ke berita dan informasi begitu mudah, tantangan utama yang timbul adalah bagaimana mengelompokkan dan mengkategorikan berita dengan cepat dan akurat. Dengan produksi berita yang semakin besar dan beragam, memahami topik berita menjadi sangat penting untuk penyampaian informasi yang efektif kepada masyarakat. Selain itu, permasalahan tambahan muncul dalam bentuk fake news dan informasi yang salah kaprah, yang memerlukan klasifikasi yang lebih cermat.

Dalam konteks ini, relevansi dan nilai penelitian ini sangat nyata. Keterbatasan metode klasifikasi berita yang ada serta kompleksitas topik-topik berita yang terus berkembang mendorong untuk mencari solusi yang lebih canggih. Penelitian di bidang ini sangat relevan dalam upaya meningkatkan pemahaman dan pengelolaan informasi di era digital. Dengan kemajuan teknologi, alat-alat pemrosesan bahasa alami dan pembelajaran mesin semakin penting untuk mengatasi tantangan klasifikasi berita. Solusi yang lebih canggih dapat membantu dalam mengotomatisasi proses kategorisasi berita dengan akurasi yang lebih tinggi dan kecepatan yang diperlukan.

Dalam konteks berita palsu dan informasi yang salah kaprah, penelitian ini bisa menjadi landasan untuk pengembangan sistem deteksi berita palsu yang lebih canggih. Ketepatan klasifikasi berita juga memiliki dampak penting pada praktik jurnalisme dan tanggung jawab media, Ini membantu menghindari bias dalam penyajian informasi dan memastikan bahwa masyarakat menerima berita yang benar-benar informatif dan objektif.

Studi-studi sebelumnya telah banyak mengkaji masalah klasifikasi berita, namun masih ada ruang untuk penelitian lebih lanjut. Beberapa penelitian sebelumnya mungkin telah mengusulkan pendekatan klasifikasi berita berdasarkan aturan dan metode tertentu, berdasarkan penelitian terdahulu yang menggunakan metode naive bayes [1]

Algoritma Naïve Bayes Classifier yang digunakan sebagai metode untuk melakukan klasifikasi data terutama untuk kebutuhan deteksi terhadap berita berita palsu/fakta.Tahapan yang telah dilakukan dimulai dengan pengumpulan data, proses tokenisasi, fase pemodelan, fase evaluasi, hingga fase deployment.[2]

Penelitian ini membahas pengelompokan atau pengklasifikasian dokumen berita secara otomatis, dikarenakan pengelompokan secara manual dengan menggunakan bantuan manusia itu tidak efisien. [3]

Penelitian ini dilakukan untuk menerapkan algoritma KNN (K-Nearest Neighbor) dalam melakukan sentimen analisis pengguna Twitter tentang isu terkait kebijakan pemerintah tentang Pembelajaran Daring. penelitian menggunakan data Tweet sebanyak 1825 data tweet Bahasa Indonesia data dikumpulkan sejak tanggal 1 Februari 2020 sampai dengan 30 September 2020. Tetapi mereka mungkin belum menyentuh pentingnya membandingkan dan menggabungkan metode yang berbeda. Selain itu, penelitian-penelitian terdahulu mungkin tidak secara khusus menerapkan Algoritma Naïve Bayes Classification dan K-Nearest Neighbors dalam konteks klasifikasi topik berita. Oleh karena itu, penelitian ini akan mengisi kesenjangan pengetahuan ini dengan eksplorasi yang lebih mendalam terhadap dua metode tersebut.

Tujuan utama dari penelitian ini adalah untuk menguji dan membandingkan efektivitas Algoritma Naïve Bayes Classification dan K-Nearest Neighbors dalam klasifikasi topik berita. Dengan demikian, Penelitian ini bertujuan untuk menyediakan pemahaman yang lebih komprehensif terkait potensi penerapan kedua metode tersebut dalam konteks yang relevan. Signifikansi penelitian ini terletak pada kontribusinya terhadap pemecahan masalah klasifikasi berita yang lebih akurat dan efisien di tengah eksponensialnya produksi berita digital. Selain itu, hasil penelitian ini dapat memiliki implikasi praktis dalam mengembangkan alat-alat otomatis untuk mengkategorikan berita, yang dapat digunakan oleh industri media, lembaga pemerintah, dan masyarakat umum. Dengan pemahaman yang lebih baik tentang efektivitas keduanya, penelitian ini dapat membantu memandu pengembangan solusi klasifikasi berita yang lebih canggih.

Untuk mencapai tujuan penelitian, peneliti akan mengintegrasikan Algoritma Naïve Bayes Classification dan K-Nearest Neighbors dalam proses klasifikasi topik berita. Algoritma Naive Bayes merupakan metode Klasifikasi sederhana. Metode ini memanfaatkan teorema probabilitas yaitu mencari peluang terbaik, dengan memprediksi probabilitas di masa depan berdasarkan informasi di masa sebelumnya. Algoritma K-NN merupakan suatu teknik yang digunakan untuk melakukan klasifikasi terhadap objek-objek berdasarkan data pembelajaran yang memiliki jarak terdekat dengan objek yang sedang diklasifikasikan. Data pembelajaran ini diwakili dalam ruang berdimensi tinggi, di mana setiap dimensi merepresentasikan ciri-ciri atau fitur-fitur yang terdapat pada data tersebut. Kami akan mengumpulkan dataset yang mencakup berbagai topik berita dan melaksanakan eksperimen dengan variasi parameter dan teknik pendekatan yang relevan dengan Informatika. Kami juga akan memanfaatkan analisis data untuk mengevaluasi kinerja kedua metode ini dalam konteks klasifikasi berita.

Hasil penelitian ini memiliki potensi untuk memperbaiki kemampuan klasifikasi topik berita secara otomatis. Jika berhasil, hasil penelitian ini dapat digunakan oleh industri media untuk menyusun berita dengan lebih efisien dan akurat, serta oleh lembaga pemerintah dalam mengawasi penyebaran informasi yang salah kaprah. Selain manfaat langsung bagi industri media dan lembaga pemerintah, hasil penelitian ini juga memiliki potensi untuk memberikan manfaat yang luas bagi masyarakat secara umum.Dengan peningkatan kemampuan klasifikasi topik berita, masyarakat akan lebih mudah untuk mengakses berita yang relevan dan terpercaya. Hal ini dapat meningkatkan literasi informasi dan membantu individu dalam membuat keputusan yang lebih informatif dan cerdas. Selain itu, temuan ini dapat menginspirasi penelitian lanjutan di bidang Informatika, terutama dalam pengembangan algoritma klasifikasi yang lebih canggih dan efektif. Implikasi praktis dari penelitian ini juga dapat membantu meningkatkan kualitas berita yang diterima oleh masyarakat, mendukung demokrasi informasi yang sehat, dan mengurangi dampak dari disinformasi digital.

2. TINJAUAN PUSTAKA

2.1. Penelitian terdahulu

Penelitian yang dilakukan oleh [4] Dengan berjudul "Perbandingan Kinerja Metode Naive Bayes Dan K-Nearest Neighbor Untuk Klasifikasi Artikel Berbahasa Indonesia". Penelitian ini membahas tentang perbandingan metode Naive Bayes dan K-Nearest Neighbor untuk mengklasifikasikan artikel jurnal berbahasa Indonesia diketahui bahwa kinerja dari metode Naive Bayes lebih unggul dari metode K-Nearest Neighbor. Terbukti bahwa dari 40 data uji yang digunakan metode Naive Bayes mampu mengklasifikasikan artikel jurnal berbahasa Indonesia sebanyak 28 dokumen. Sedangkan untuk metode K-Nearest Neighbor dari 40 data uji metode ini hanya dapat mengklasifikasikan artikel jurnal berbahasa Indonesia sebanyak 16 dokumen. Hal tersebut dapat dipengaruhi jumlah data yang digunakan dan tahapan preprocessing yang dilakukan. Oleh karena itu,

disarankan untuk menambah data set dan melengkapi tahapan preprocessing seperti melakukan stemming kata pada penelitian selanjutnya.

Penelitian yang dilakukan oleh [5] Dengan Berjudul "Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity" Penelitian ini membahas klasifikasi berita online dengan menggunakan tf-idf dan cosine similarity, memerlukan proses preprocessing yaitu tokenizing, stopword dan stemming dapat memperkecil term sehingga mempercepat proses perhitungan pembobotan term menggunakan tf-idf dan mempercepat proses cosine similarity. Tujuannya adalah untuk mempermudahkan human error serta mengurangi terjadinya kesalahan pengkategorian. klasifikasi mampu mengelompokkan berita dengan tingkat akurasi sebesar 91.25%.

Penelitian yang dilakukan oleh [6]Dengan berjudul "Klasifikasi Artikel Berita Secara Otomatis Menggunakan Metode Naive Bayes Classifier Yang Dimodifikasi" Penelitian ini memaparkan modifikasi metode Naïve Bayes Classifier dengan melakukan pembobotan kata berdasarkan posisinya dalam berita. Percobaan dilakukan pada 900 dokumen berita. Sembilan ratus dokumen tersebut dibagi menjadi 9 kategori, sehingga masing-masing kategori diujikan 100 dokumen. Untuk mengetahui pengaruh jumlah data latih terhadap efektifitas naïve bayes classifier maka diambil beberapa kombinasi banyaknya dokumen latih dan dokumen uji.

Penelitian yang dilakukan oleh [7]Dengan Berjudul "Klasifikasi berita menggunakan algoritma SVM" Penelitian ini membahas tentang pengelompokan sebuah berita menjadi beberapa kategori seperti ekonomi, olahraga, politik dll. Pada penelitian ini menggunakan algoritma support vector machine dalam melakukan proses klasifikasi teks.

Penelitian yang dilakukan oleh [1]Dengan berjudul "Implementasi Algoritma Naïve Bayes Classifier untuk mendeteksi berita palsu pada media sosial" Penelitian ini membahas Algoritma Naïve Bayes Classifier yang digunakan sebagai metode untuk melakukan klasifikasi data terutama untuk kebutuhan deteksi terhadap berita berita palsu/fakta.Tahapan yang telah dilakukan dimulai dengan pengumpulan data, proses tokenisasi, fase pemodelan, fase evaluasi, hingga fase deployment.

Penelitian yang dilakukan oleh [8]Dengan berjudul "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter" Dalam penelitian ini sebanyak 2000 dataset text posting user dianalisis yang bersumber dari aplikasi media sosial Twitter. Penelitian ini menggunakan turunan dari algoritma Naïve Bayes yaitu Multinomial Naïve Bayes untuk mengoptimalkan hasil klasifikasi. Tiga label kelas yang digunakan untuk mengklasifikasi sentiment masyarakat yaitu sentimen positif, negative dan netral. Tahapanya dimulai dengan text preprocessing; cleaning, case folding, tokenisasi, filtering dan stemming. Kemudian dilanjutkan dengan pembobotan menggunakan pendekatan TF-IDF. Untuk mengevaluasi hasil klasifikasi, data diuji menggunakan confusion matrix dengan menguji akurasi, precision dan recall.

Penelitian yang dilakukan oleh [9]Dengan berjudul "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation" Penelitian ini bertujuan untuk melakukan klasifikasi opini-opini wisatawan menjadi dua kelas yaitu positif dan negatif serta melakukan pemodelan topik pada kedua kelas tersebut. Pemodelan topik bertujuan untuk mengetahui topik yang sering dibicarakan pada masing-masing kelas. Tahapan dari penelitian ini meliputi pengumpulan data, pembersihan data, transformasi data, klasifikasi data dengan metode Naive Bayes dan penggunaan metode Latent Dirichlet Allocation (LDA) untuk pemodelan topik.

Penelitian yang dilakukan oleh [10]yang berjudul "Filtering Spam Email Menggunakan Metode Naive Bayes" Naive Bayes merupakan metode Klasifikasi sederhana. Metode ini memanfaatkan teorema probabilitas yaitu mencari peluang terbaik, dengan memprediksi probabilitas di masa depan berdasarkan informasi di masa sebelumnya. Tujuan utama dalam penulisan skripsi ini adalah mengkaji penerapan metode Naive Bayes untuk menentukan email spam dan email ham. Hasil pengujian aplikasi terhadap 5 email yang terdiri dari 2 email spam dan 3 email ham.

Penelitian yang dilakukan oleh [11]dengan berjudul "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid" penelitian ini dilakukan sebagai langkah antisipasi terhadap pandemi Covid-19 dengan memprediksi tingkat penyebaran Covid-19 terutama di Indonesia. Metode penelitian yang di terapkan pada penelitian ini ialah analisis masalah dan studi literatur, mengumpulkan data dan implementasi. Penerapan metode naive bayes diharapkan mampu memprediksi tingkat penyebaran COVID-19 di Indonesia.

Penelitian yang dilakukan oleh [12]dengan berjudul "Prediksi Tingkat Kepuasan dalam Pembelajaran Daring Menggunakan Algoritma Naïve Bayes" Riset ini bertujuan buat memprediksi tingkatan kepuasan mahasiswa dalam pendidikan daring serta membagikan donasi pada akademi besar dalam mengambil kebijakan yang berhubungan dengan kenaikan mutu pendidikan secara daring. Informasi yang digunakan diperoleh dengan membagikan kuesioner kepada mahasiswaangkatan 2020/2021 sebanyak 110 mahasiswa. Parameter yang ada pada kuesioner ialah komunikasi dosen, atmosfer pendidikan daring, evaluasi terhadap mahasiswa, penyampaian modul. Naïve Bayes ialah salah satu metode prediksi buat mencari probabilitas simple bersumber pada teorema bayes dengan anggapan independensi yang kokoh.

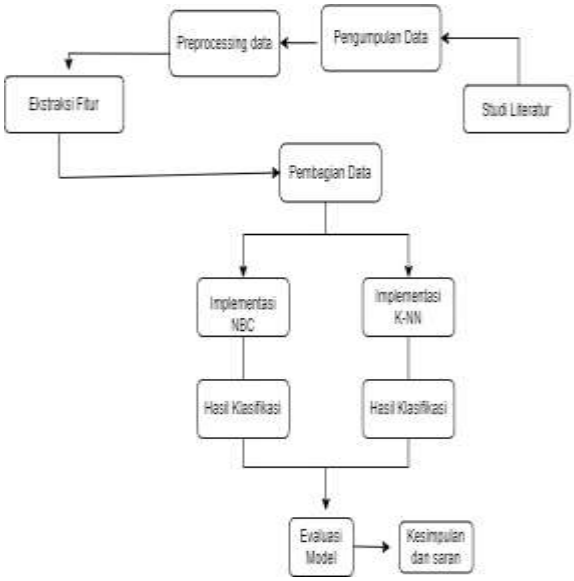
Penelitian yang dilakukan oleh[13]dengan berjudul "Analisis performa metode K-NN pada

Dataset pasien pengidap Kanker Payudara" Penelitian ini mencoba menerapkan metode KNN pada dataset pasien pengidap penyakit kanker payudara, k yg diterapkan adalah k=3 hingga k=5 serta menerapkan crossvalidation dengan kfold=5, setelah dilakukan pengujian maka dengan metode KNN diperoleh hasil tertinggi untuk Akurasi dengan nilai 0,93 pada 20% keempat (K3), 20% Pertama(K4) dan 20% pertama(K5), untuk Presisi dengan nilai 0,97 pada 20% keempat(K3), untuk Recall dengan nilai 0,98 pada 20% ketiga (K3) dan F-measure dengan nilai 0,94 pada 20% keempat(K3) dan 20% ketiga(K5).

Penelitian yang dilakukan oleh [14] dengan berjudul "Perbandingan Metode Naïve Bayes, Knn, Decision Tree Pada Laporan Water Level Jakarta" Penelitian ini membahas tentang perbandingan antara metode Naivebayes, KNN, Decision Tree. Dimana data dari penelitan adalah data set laporan ketinggian air di Jakarta berasal dari data.go.id, pada penelitian ini akan diukur confusion matrix, precision, recall, accuracy, hingga f-measure kemudian juga dihitung root mean squere error dari tiap-tiap metode, dari perhitungan tersebut metode Decission tree mendapatkan accuracy tertinggi hingga 96.56% sehingga dapat disimpulkan metode klasifikasi decision tree lebih baik dari metode Naivebayes maupun KNN.

3. METODE PENELITIAN

Penerapan Algoritma Naïve Bayes Classification dan K-Nearest Neighbors dalam konteks klasifikasi topik berita berjumlah 9 tahapan. Tahapan tersebut adalah Pengumpulan Data, Pengambilan Sampel, Pengelompokan data berdasarkan kategori, Preprocessing Data, Ekstraksi Fitur, Pembagian Data, Penerapan Algoritma Naïve Bayes Classification, Penerapan K-Nearest Neighbors (K-NN), Evaluasi Model.

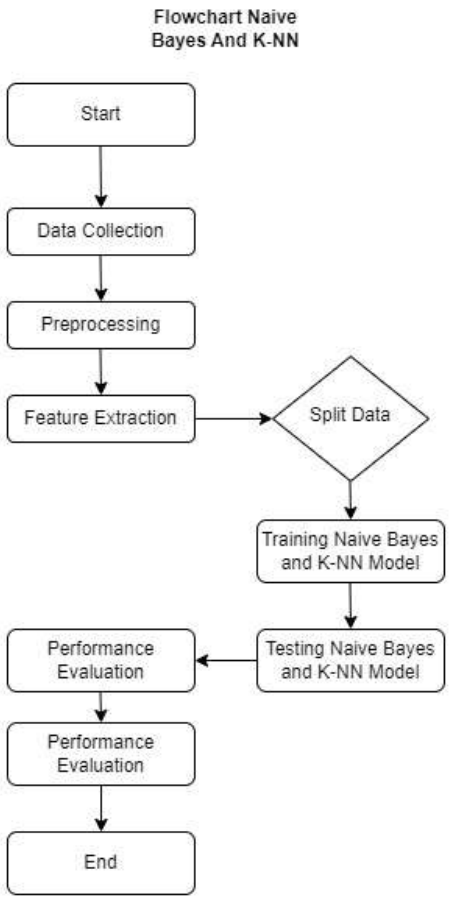


Gambar 1. flowchart metode penelitian

Dari Gambar 1 tentang metode penelitian dapat dijelaskan sebagai berikut;

a. Studi Literatur Tahap ini bertujuan untuk memahami teori dasar di balik kedua algoritma dan konsep-konsep yang terkait dengan klasifikasi teks dan topik berita. Studi literatur ini akan membantu merumuskan pemahaman awal dan kerangka kerja penelitian.
b. Pengumpulan data adalah langkah awal dalam penelitian. Data ini akan digunakan sebagai bahan mentah untuk melatih dan menguji model klasifikasi.
c. Preprocessing data adalah langkah penting untuk memastikan data siap digunakan dalam model klasifikasi. Data yang sudah bersih dan terstruktur akan menghasilkan hasil yang lebih baik.
d. Ekstraksi fitur adalah Fitur-fitur yang digunakan untuk mewakili teks berita dalam bentuk numerik yang dapat digunakan oleh algoritma klasifikasi. Pemilihan metode ekstraksi fitur yang tepat akan memengaruhi kinerja model.
e. Pembagian data adalah Membagi data menjadi set pelatihan (Training set dan Testing Set) Pembagian data menjadi data training dan testing adalah salah satu praktik terpenting dalam pembelajaran mesin, karena membantu memastikan bahwa model yang dikembangkan mampu melakukan prediksi yang akurat dan dapat menggeneralisasi pada data baru yang belum pernah dilihat sebelumnya.
f. Penerapan Algoritma Naïve Bayes Classification, Selama tahap ini, model statistik Naïve Bayes akan digunakan untuk memprediksi kelas topik berita berdasarkan fitur-fitur yang diekstrak. Model akan disesuaikan dengan data pelatihan.
g. Penerapan K-Nearest Neighbors (K-NN)
h. Model K-NN akan digunakan untuk melakukan klasifikasi berdasarkan kesamaan antara data pengujian dan data pelatihan. Pemilihan parameter K akan menjadi bagian dari eksperimen.
i. Evaluasi model adalah langkah penting untuk menilai sejauh mana model dapat melakukan klasifikasi dengan baik. Metrik ini membantu membandingkan kedua algoritma dan menentukan yang lebih efektif dalam konteks klasifikasi berita.
j. Kesimpulan dan Saran
   Laporan penelitian adalah dokumen akhir yang mendokumentasikan seluruh penelitian dan dapat digunakan untuk berbagi pengetahuan dengan komunitas ilmiah atau praktisi lainnya.

Baik algoritma Naive Bayes maupun KNN diterapkan untuk mengklasifikasikan topik berita di detik.com. Naive Bayes menghitung probabilitas berdasarkan asumsi independensi antar fitur, sementara KNN mengklasifikasikan instance berdasarkan kelas tetangga terdekatnya. Kinerja kedua algoritma dievaluasi menggunakan metrik klasifikasi standar.



Gambar 2. Flowchart Naive bayes dan K-NN

3.1. Sumber Data

Data yang digunakan dalam penelitian ini diperoleh melalui teknik web scraping dari situs web Detik.com (https://www.detik.com). Kami menjelaskan proses pengumpulan data, jenis data yang diperoleh, serta langkah-langkah yang diambil untuk memastikan kualitas dan ketersediaan data.

3.2. Populasi dan Sampel

Populasi penelitian adalah kelompok atau keseluruhan yang merupakan subjek atau objek penelitian yang relevan dengan topik penelitian ini. Dalam konteks penelitian ini, populasi penelitian adalah semua artikel berita yang terdapat di situs web Detik.com (https://www.detik.com/). Artikel berita ini mencakup berbagai topik dan kategori, termasuk politik, bisnis, teknologi, hiburan, dan lainnya. Karena populasi ini mencakup seluruh isi situs web, kami akan menggunakan sampel yang representatif untuk mewakili bagian dari populasi ini.

a. Populasi
   Jumlah berita di situs tersebut dalam satu sampai dua bulan terakhir pada tahun 2023 yaitu pada bulan november sampai desember terdapat 17.718 data berita.
b. Pemilihan Sampel
   Pengambilan Sampel dilakukan dalam 5 kategori yaitu politik, ekonomi, bisnis, pendidikan dan olahraga peneliti menggunakan metode pengambilan sampel berdasarkan 5 kategori tersebut untuk memilih artikel berita dari situs web Detik.com. Dalam metode ini, setiap artikel berita dalam populasi memiliki probabilitas yang sama untuk dipilih sebagai bagian dari sampel. Pengelompokan Berdasarkan Kategori Sampel akan dibagi ke dalam kelompok berdasarkan kategori berita, seperti politik, ekonomi, bisnis, pendidikan dan olahraga. Ini akan memastikan bahwa setiap kategori berita terwakili dalam sampel.
c. Jumlah Sampel
   Jumlah sampel yang akan diambil akan ditentukan berdasarkan pertimbangan statistik yang memungkinkan untuk mencapai tingkat signifikansi yang diperlukan dalam analisis berdasarkan algoritma Naive Bayes Classification dan K-NN. Karakteristik Sampel Setiap artikel berita yang terpilih dalam sampel akan dicatat untuk data analisis. Data yang akan diambil dari setiap artikel berita termasuk judul, teks berita, kategori berita, dan atribut lain yang relevan. Sampel dari penelitian ini akan dipilih secara acak dari populasi yang ada. Jumlah sampel yang akan diambil akan bergantung pada besarnya populasi dan tingkat ketepatan yang diinginkan. Penelitian ini memutuskan untuk mengambil sampel sebanyak 39,41% dari populasi, dan pada bulan november sampai desember 2023 terdapat 17.718 berita yang sesuai 5 kategori yang di cari pada situs Detik.com, jumlah sampel yang diambil adalah 39,41% dari 17.718, yaitu 6.999 berita dalam periode 2 bulan terakhir pada bulan November sampai Desember 2023 dengan tahapan pengambilan data di ambil dalam beberapa tahap pengambilan.

3.3. Teknik Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan teks berita yang diperoleh dari berbagai kategori di situs web Detik.com. observasi atau pengumpulan data dilakukan secara berkala sesuai dengan jadwal atau interval waktu yang ditetapkan. Waktu observasi dapat bervariasi tergantung pada tujuan penelitian dan sumber daya yang tersedia. pendekatan yang digunakan dalam pengumpulan data berita dari situs web Detik.com ini menggunakan metode observasi Pengamatan terjadwal. Dalam metode ini, peneliti mungkin menentukan jadwal atau interval waktu tertentu untuk mengumpulkan data. Contohnya, penelitian ini dapat memutuskan untuk mengumpulkan data berita setiap hari pada pukul 00.00 waktu setempat, atau mungkin sekali dalam seminggu pada hari tertentu. Pendekatan ini dapat lebih terstruktur dan memungkinkan peneliti untuk mengontrol waktu pengumpulan data.

3.4. Teknik Analisis Data

Data yang digunakan dalam penelitian ini merupakan teks berita yang diperoleh dari berbagai

kategori di situs web berita Detik.com adapun tahapan tahapan nya adalah sebagai berikut.

a. Preprocessing data
   Data teks berita dapat mengandung karakteristik seperti tanda baca, angka, dan karakter khusus yang tidak relevan. Oleh karena itu, kami akan melakukan pembersihan data dengan menghapus karakter yang tidak diperlukan.
b. Pembentukan Matriks TF-IDF
   Pembentukan Matriks menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) untuk mengukur pentingnya kata-kata dalam setiap artikel berita. Ini akan menghasilkan matriks TF-IDF yang akan digunakan sebagai representasi data dalam analisis.
c. Analisis Data
   Setelah persiapan data selesai, tahap analisis data akan dimulai. Dalam penelitian ini akan menerapkan dua metode analisis utama.
d. Naive Bayes Classification
   Algoritma Naive Bayes Classification akan digunakan untuk mengklasifikasikan topik berita berdasarkan matriks TF-IDF yang telah dibentuk. Metode ini akan menghasilkan prediksi klasifikasi topik berita.
e. K-NN (K-Nearest Neighbors)
   Metode K-NN akan digunakan sebagai pendekatan alternatif untuk klasifikasi topik berita. Dalam metode ini, akan dilakukan identifikasi tetangga terdekat dari setiap artikel berita dalam ruang fitur TF-IDF.
f. Evaluasi
   Evaluasi hasil analisis adalah tahap penting dalam penelitian ini. Peneliti akan menggunakan metrik evaluasi yang relevan.

3.4.1. Naive Bayes Classification:

Naive Bayes Classification (NBC) merupakan sebuah metode pengklasifikasi probabilistik yang sederhana. Metode ini menghitung sejumlah probabilitas dengan mendasarkan pada pengamatan frekuensi dan kombinasi nilai-nilai dalam dataset yang diberikan. Disini peneliti menggunakan tipe Algoritma Multinomial Naive Bayes yang merupakan salah satu metode pembelajaran probabilistik didasarkan pada teorema Bayes yang digunakan dalam Natural Language Processing (NLP). Algoritma ini bekerja pada konsep term frequency yang berarti berapa kali kata tersebut muncul dalam sebuah dokumen. Model ini menjelaskan dua fakta yaitu apakah kata tersebut muncul dalam sebuah dokumen atau tidak serta frekuensinya kemunculan dalam dokumen. Multinomial Naïve Bayes dapat diformulasikan sebagai berikut, rumus 1.

$$p(p|n) \propto P(p) \prod_{1 \le k \le nd} P\,(t_k\,|p)$$

dimana P(t_k|p): probabilitas munculnya dokumen text (t_k), n adalah jumlah dokumen dan p adalah polaritas. Kemudian untuk menghitung polaritasnya atau dokumen yang mempunyai kemiripan dirumuskan sebagai berikut, rumus 2

$$p(t_k\,|p) = \frac{count(t_k|p)+_1}{count\left(t_p\right) + |v|}$$

Dimana (t_k| p) adalah jumlah t_k muncul di dokumen text yang memiliki polaritas p dan jumlah (t_p) berarti jumlah token yang ada di artikel berita dengan polaritas p.

3.4.2. K- Nearest Neighbors

Algoritma K-Nearest Neighbor (KNN) digunakan untuk mengklasifikasikan objek baru berdasarkan kesamaannya dengan data pembelajaran yang telah ada. Teknik ini memilih kelas atau label untuk objek baru berdasarkan mayoritas kelas dari k tetangga terdekatnya dalam data pembelajaran. Jumlah tetangga yang dipertimbangkan ini ditentukan oleh parameter k.

$$d\left(x_i, x_j\right) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

Dimana d (x_i,x_j) adalah jarak euclidean, x_i adalah record ke i, x_j record ke j dan a_r data ke r. 2) Urutkan dengan nilai jarak Euclidean, 3) Menentukan k record klasifikasi terdekat, dan 4) Target outputnya adalah kelas mayoritas.

Untuk analisis perbandingan, penelitian ini dapat mengukur kinerja kedua metode menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Hasil evaluasi ini akan memberikan pemahaman yang lebih mendalam tentang sejauh mana setiap metode efektif dalam mengklasifikasikan topik berita. Terlebih lagi, penelitian ini juga dapat mengeksplorasi apakah kombinasi atau ensemble dari kedua metode ini dapat meningkatkan kinerja klasifikasi.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Pengumpulan Data

Dataset yang digunakan sebanyak 6.999 topik judul berita dalam periode 2 bulan terakhir, topik berita yang di ambil ada beberapa kategori seperti politik, ekonomi, bisnis, pendidikan dan olahraga masing masing kategori diambil secara acak sesuai perilisan berita tersebut selama 2 bulan terakhir. Metode pengumpulan ini yakni menarik data di internet yang bersumber dari website detik.com dengan cara web scraping. Data yang sudah ditarik dimasukkan kedalam excel kemudian dimasukkan ke dalam database. Dibawah ini merupakan contoh data yang diambil dari internet lalu diinput di excel yang dapat dilihat pada gambar :

Gambar 3. Hasil Scraping

Preprocessing dilakukan agar data dapat diproses ke tahap klasifikasi. Adapun tahapan preprocessing yang dapat dilihat pada gambar 3 beberapa hasil tokenisasi dan perhitungan TF-IDF

Gambar 4. Term hasil Tokenizing

4.2. Hasil Ekstraksi Fitur

1. Hasil Perhitungan Term Frequency(TF)

Gambar 5. Hasil Perhitungan Term Frequency

2. Hasil Perhitungan Invers Document Frequency

Gambar 6. Hasil Perhitungan IDF

3. Hasil Perhitungan TF-IDF (dalam skala 0-1)

Gambar 7. Hasil perhitungan TF-IDF

4.3. Hasil Pemisahan Data

Gambar 8 Hasil Pemisahan Data

Gambar 9. x_test dan x_train

Gambar 10. y_train dan y_test

4.4. Hasil Implementasi Algoritma Naive Bayes dan K-NN

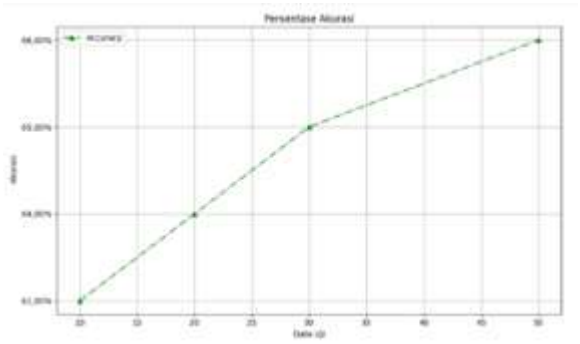
4.4.1. Klasifikasi Naive Bayes

Pada klasifikasi jumlah data yang digunakan sebesar 6.999 data, kemudian akan dilakukan terlebih

dahulu pembagian data training dan data uji dimana akan dibagi menjadi beberapa bagian, kemudian pada data uji akan di test untuk implementasi algoritma NB dalam mencari akurasi. Pembagian data tersebut antara lain data training 70% data uji 30%, data training 80% data uji 20%, data training 50% data uji 50% dan Data training 90% Data uji 10% dengan akurasi yang paling tinggi di dapat sebesar 66,00%. Berikut data-data yang sudah di test dapat dilihat pada tabel 1 :

Tabel 1. Hasil Akurasi Naive Bayes

Data training	Data uji	Akurasi
80%(5.599 data)	20%(1.400 data)	65,00%
70%(4.899 data)	30%(2.100 data)	64,00%
50%(3.499 data)	50%(3.500 data)	61,00%
90%(6.299 data)	10%(700 data)	66,00%



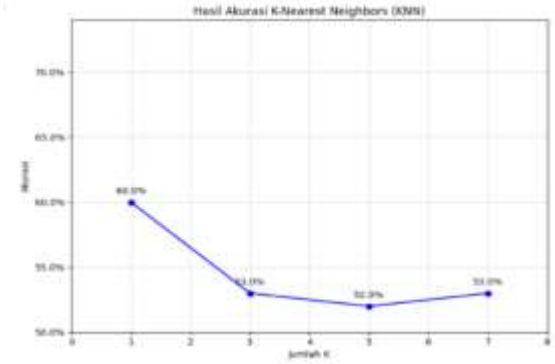
Gambar 11. Diagram Hasil Akurasi Naive Bayes

4.4.2. Klasifikasi K-Nearest Neighbors

Total data yang dipakai pada klasifikasi ini adalah 6.999. Kemudian dilakukan pemecahan data training dan data uji yang dipecah menjadi beberapa bagian, kemudian pada data uji akan di test untuk implementasi algoritma KNearest Neighbor dalam mencari akurasi. Pembagian data tersebut antara lain data training 70% data uji 30%, data training 80% data uji 20%, dan data training 50% data uji 50%. Dan pada jumlah (K) nya akan ditentukan antara lain 1, 3, 5, dan 7.

a. Klasifikasi K-Nearest Neighbor dengan 50% data training dan 50%

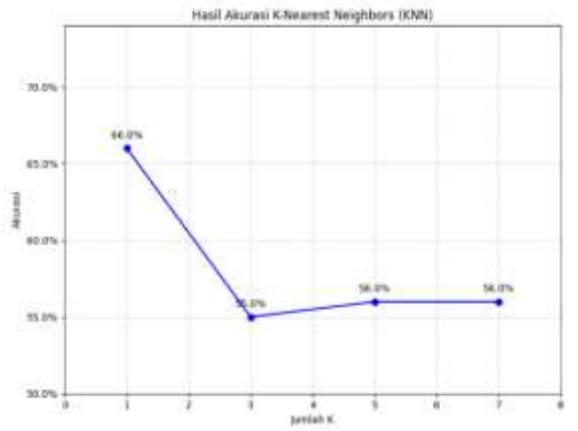
Data uji Pengujian algoritma K-Nearest Neighbor menggunakan data uji 50% didapatkan akurasi terbaik sebesar 60,00% pada k = 1. Berikut hasil akurasi data uji 50% dapat dilihat pada gambar 12:



Gambar 12. Akurasi Data Uji 50%

b. Klasifikasi K-Nearest Neighbor dengan 70% data training dan 30% data uji

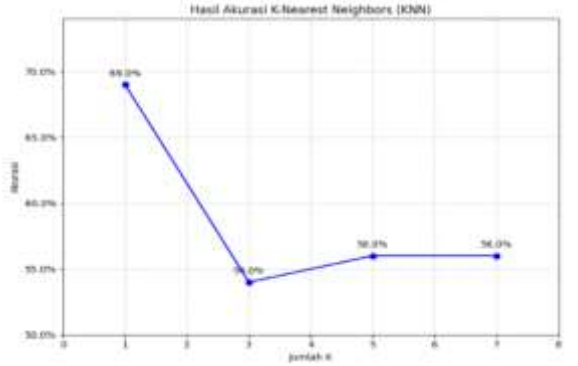
Pengujian algoritma K-NN menggunakan data uji 10% didapatkan akurasi terbaik pada k= 1 dengan akurasi sebesar 66,00%. Berikut hasil akurasi data uji 10% dapat dilihat pada gambar 13.



Gambar 13. Akurasi Data Uji 30%

c. Klasifikasi K-Nearest Neighbor dengan 80% data training dan 20% data uji

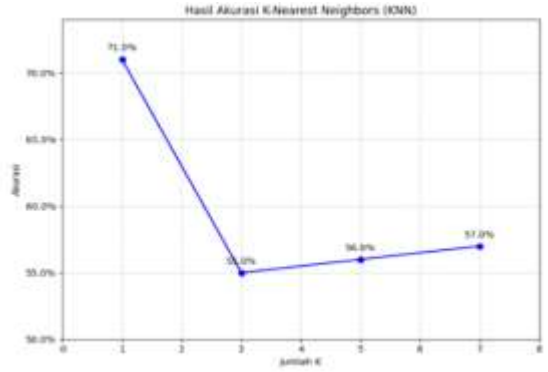
Pengujian algoritma K-NN menggunakan data uji 20% didapatkan akurasi terbaik pada k=1 dengan akurasi sebesar 69,00%. Berikut hasil akurasi data uji 20% dapat dilihat pada gambar 14 :



Gambar 14. Akurasi Data Uji 20%

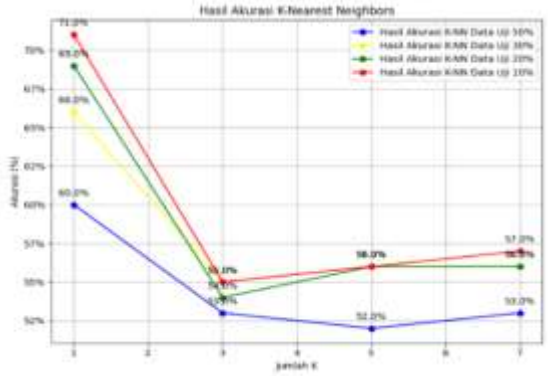
d. Klasifikasi K-Nearest Neighbor dengan 90% data training dan 10% data uji

Pengujian K-NN pada data uji 10% didapatkan akurasi terbaik sebesar 71,00% pada k = 1. Berikut hasil akurasi data uji 10% dapat dilihat pada gambar 15.



Gambar 15. Data Uji 10%

Pengujian algoritma K-Nearest Neighbors menggunakan beberapa bagian data uji dengan beberapa jumlah k didapatan akurasi terbaik pada jumlah k masing masing dari data uji tersebut.Berikut yaitu hasil semua data testing K-Nearest Neighbor yang dapat dilihat pada tabel 6 :



Gambar 16. Hasil Akurasi K-NN

4.5. Hasil Perbandingan Klasifikasi Topik Berita

Perbandingan Hasil Naïve Bayes dan K-Nearest Neighbors Ketika akurasi dari K-NN dan NB dibandingkan, ditemukan akurasi K-NN kian unggul dari NB. Karena pada akurasi pengujian K-NN menghasilkan akurasi tertinggi yaitu 71,00% dengan data uji 10% pada k = 1 jika dibandingkan dengan akurasi NB yang menghasilkan akurasi sebesar 66,00% pada data uji 10%. Berikut hasil akurasi kedua algoritma tersebut yang dapat dilihat pada tabel 6 :

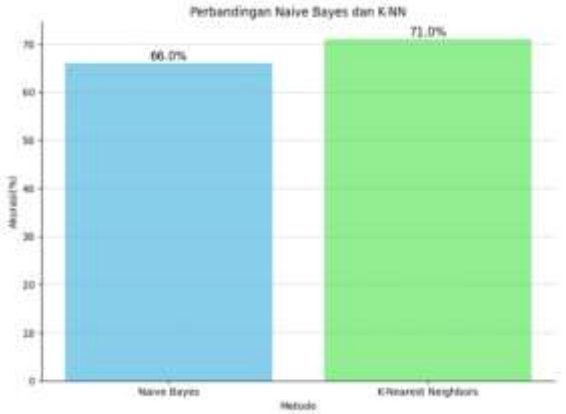
Tabel 2. Hasil Akurasi K-NN

Data Training	Data Uji	Jumlah K	Akurasi
50%	50%	1	**60,00%**
		3	53,00%
		5	52,00%
		7	53,00%
70%	30%	1	**66,00%**
		3	55,00%
		5	56,00%
		7	56,00%
80%	20%	1	**69,00%**
		3	54,00%
		5	56,00%
		7	56,00%
90%	10%	1	**71,00%**
		3	55,00%
		5	56,00%
		7	57,00%

4.6. Hasil Perbandingan Akurasi Naive Bayes Dan K-Nearest Neighbors

Tabel 3. Hasil Perbandingan

Metode	Akurasi
Naive bayes	66.00%
K-Nearest Neighbors	71.00%



Gambar 17. Hasil Perbandingan

Dalam konteks klasifikasi topik berita, kedua metode ini dapat memberikan hasil yang baik tergantung pada kompleksitas dan karakteristik dari dataset berita yang digunakan. Naïve Bayes cenderung efektif dalam kondisi di mana independensi antar atribut terpenuhi, sementara K-NN dapat unggul dalam situasi di mana terdapat pola spasial yang jelas dalam data berita. pada penelitian ini penggunaan data berita yang dipakai sebanyak 6.999 data berita yang dibagi 5 kategori yaitu politik, ekonomi, bisnis, pendidikan dan olahraga. Dari 6.999 data berita kategori politik memiliki data sebesar 1.524 data, untuk kategori ekonomi memiliki 1.521, untuk kategori bisnis memiliki 1.512 data, untuk kategori pendidikan 1.088 data dan untuk kategori olahraga memiliki data sebesar 1.354 data yang akan diklasifikasikan dalam pengujian untuk metode K-Nearest Neighbor dan naive bayes, pertama untuk model K-NN dilakukan pemecahan data training dan data uji yang dipecah menjadi beberapa bagian, kemudian pada data uji akan di test untuk implementasi algoritma K- Nearest Neighbor dalam mencari akurasi.

Pembagian data tersebut antara lain Data training 90% Data uji 10%, data training 80% data uji 20%, data training 70% data uji 30%, dan data training 50% data uji 50%. Dan pada jumlah (K) nya akan ditentukan antara lain 1, 3, 5, dan 7. Untuk metode Naïve Bayes akan dilakukan terlebih dahulu pembagian data training dan data uji dimana akan dibagi menjadi beberapa bagian, kemudian pada data

uji akan di test untuk implementasi algoritma NB dalam mencari akurasi. Pembagian data tersebut antara lain Data training 90% Data uji 10%, data training 80% data uji 20%, data training 70% data uji 30%, dan data training 50% data uji 50%. Setelah mendapatkan hasil dari penelitian ini selanjutnya didapatkan akurasi dari masing- masing algoritma yang digunakan untuk di jadikan perbandingan, Dari hasil pengujian ini dapat disimpulkan bahwa pengujian analisis algoritma Naïve Bayes dan K-Nearest Neighbors ini didapatkan hasil akurasi terbaik dalam klasifikasi data yaitu pada algoritma K-Nearest Neighbors yang mendapatkan akurasi sebesar 71,00% yang didapatkan pada data uji 10% dengan k = 1 dibandingkan algoritma Naïve Bayes yang hanya mendapatkan akurasi sebesar 66,00% pada data uji 10%.

5. KESIMPULAN DAN SARAN

K-Nearest Neighbors memiliki akurasi klasifikasi yang sedikit lebih tinggi (71,00%) daripada Naive Bayes Classification, berdasarkan perbandingan. Semuanya memiliki kelebihan dan kekurangan. K-Nearest Neighbors memiliki kelebihan, seperti kemampuan untuk menangani data non-numerik dan kinerja yang dapat ditingkatkan dengan memilih nilai k yang tepat. Mereka juga membutuhkan banyak memori, sensitif terhadap outlier, dan lambat dalam memprediksi data baru.Kelebihan dari Klasifikasi Naive Bayes adalah sederhana, mudah digunakan, dan bekerja dengan baik pada dataset kecil. Namun, kekurangannya termasuk asumsi kemerdekaan fitur yang kuat, sensitif terhadap nilai fitur yang jarang terjadi, dan mengalami penurunan kinerja pada dataset besar dan kompleks.Sangat penting untuk mempertimbangkan kebutuhan dan preferensi pengguna saat memilih metode klasifikasi. Untuk penelitian selanjutnya, cakupan kedua algoritma harus diperluas.

DAFTAR PUSTAKA

[1] N. Agustina and M. Hermawati, "Implementasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Berita Palsu pada Sosial Media," *Fakt. Exacta*, vol. 14, no. 4, pp. 1979–276, 2021, doi: 10.30998/faktorexacta.v14i4.11259.

[2] A. Y. Muniar, P. Pasnur, and K. R. Lestari, "Penerapan Algoritma K-Nearest Neighbor pada Pengklasifikasian Dokumen Berita Online," *Inspir. J. Teknol. Inf. dan Komun.*, vol. 10, no. 2, p. 137, 2020, doi: 10.35585/inspir.v10i2.2570.

[3] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, p. 121, 2021, doi: 10.22146/ijccs.65176.

[4] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, "Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 427–434, 2018, doi: 10.25126/jtiik.201854773.

[5] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, "Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 1, pp. 306–312, 2018, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/796

[6] Firdaus Mahmudy and Wahyu Widodo, "Klasifikasi Artikel Berita Secara Otomatis Menggunakan," *Tekno*.

[7] R. Nanda, E. Haerani, S. K. Gusti, and S. Ramadhani, "Klasifikasi Berita Menggunakan Metode Support Vector Machine," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 2, pp. 269–278, 2022, doi: 10.32672/jnkti.v5i2.4193.

[8] Yuyun, Nurul Hidayah, and Supriadi Sahibu, "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 4, pp. 820–826, 2021, doi: 10.29207/resti.v5i4.3146.

[9] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 123–131, 2021, doi: 10.29207/resti.v5i1.2587.

[10] A. D. Wibisono, S. Dadi Rizkiono, and A. Wantoro, "Filtering Spam Email Menggunakan Metode Naive Bayes," *TELEFORTECH J. Telemat. Inf. Technol.*, vol. 1, no. 1, 2020, doi: 10.33365/tft.v1i1.685.

[11] Rayuwati, Husna Gemasih, and Irma Nizar, "IMPLEMENTASI AlGORITMA NAIVE BAYES UNTUK MEMPREDIKSI TINGKAT PENYEBARAN COVID," *Jural Ris. Rumpun Ilmu Tek.*, vol. 1, no. 1, pp. 38–46, 2022, doi: 10.55606/jurritek.v1i1.127.

[12] A. R. Damanik, S. Sumijan, and G. W. Nurcahyo, "Prediksi Tingkat Kepuasan dalam Pembelajaran Daring Menggunakan Algoritma Naïve Bayes," *J. Sistim Inf. dan Teknol.*, vol. 3, pp. 88–94, 2021, doi: 10.37034/jsisfotek.v3i3.49.

[13] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[14] D. Marutho, "Perbandingan Metode Naive Bayes , KNN , Decision Tree Pada Laporan Water Level Jakarta," *Manaj. Inform. AMIK JTC Semarang*, vol. 15, no. 2, pp. 90–97, 2019.